

Importing Data into R Studio

Using R Studio



Dataset

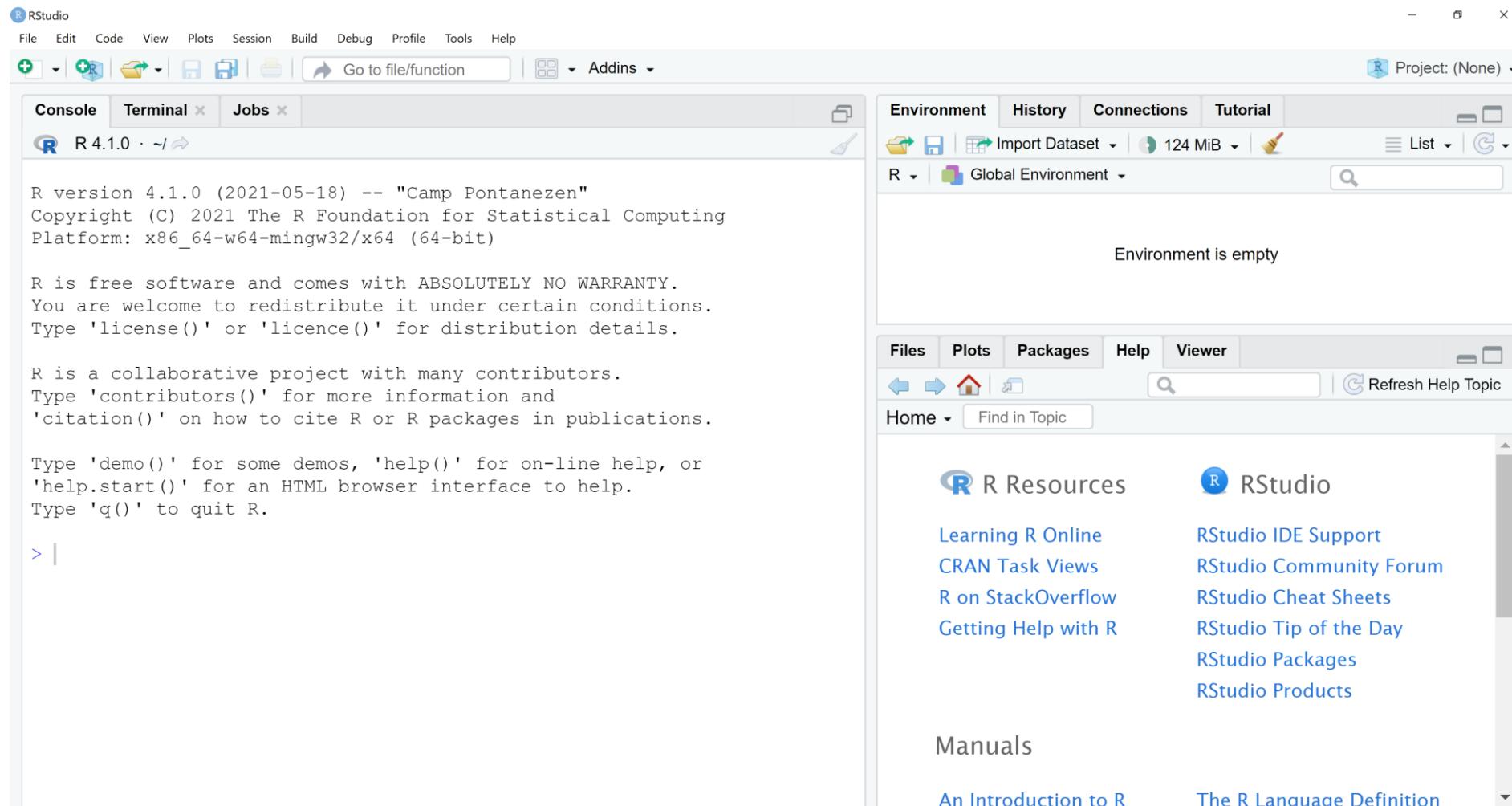
- This tutorial is a walkthrough with a sample set of data. You may use this to walk through the tutorial, if you wish, but for your assignments, you will be asked to use your own dataset (as specified within the course).

Dataset reference:

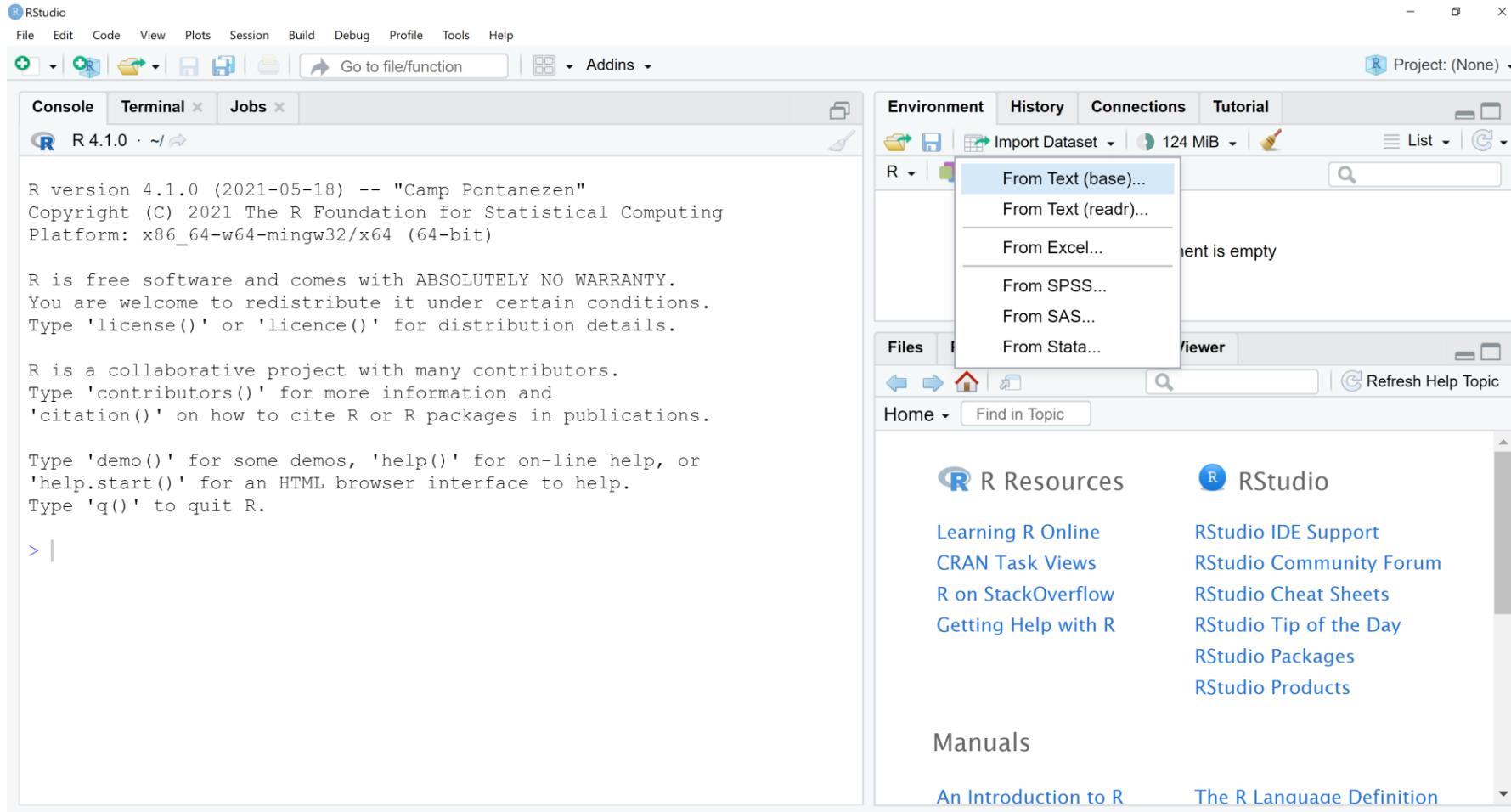
Skoryk, M. (2021). Sepsis Prediction from Clinical Data. Version 1.
Retrieved from <https://www.kaggle.com/maxskoryk/datasetsepsis>



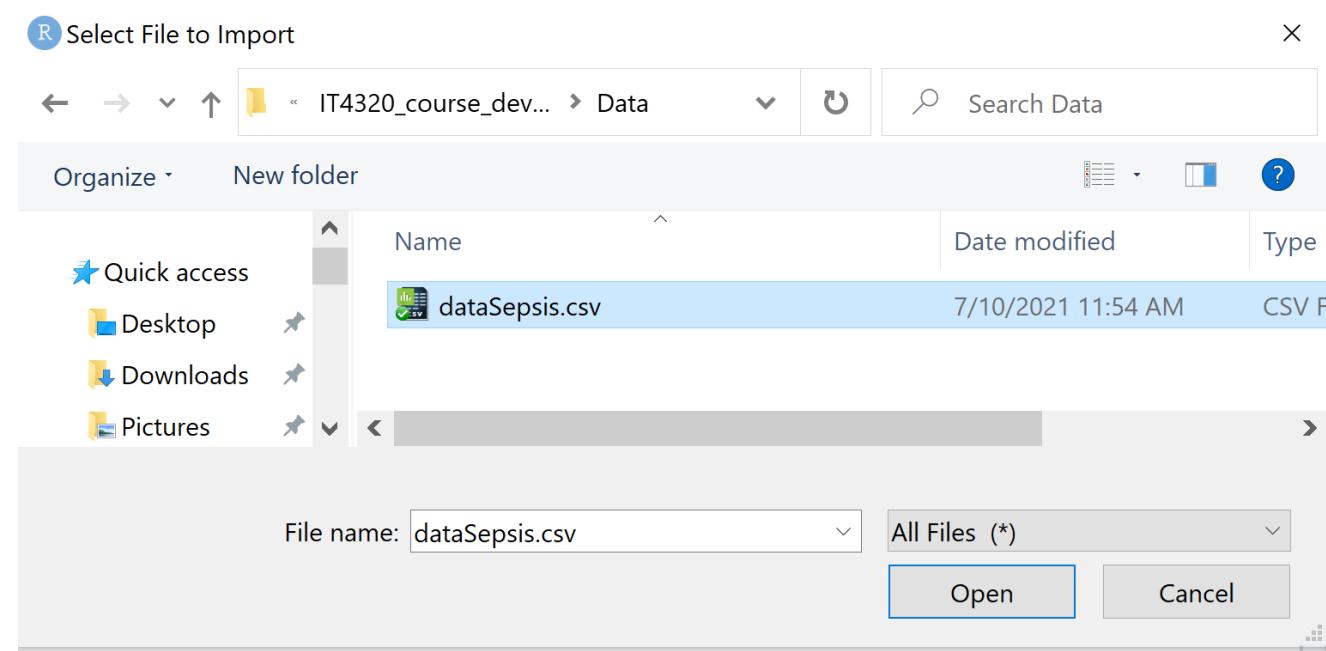
Open R Studio



Click on “Import Data” and Choose “From Text (base)”



Navigate to Your Dataset, Then Click “Open”



Select Options to Import Your Data Based on the Format of Your Text File

The screenshot shows the RStudio interface with the 'Import Dataset' dialog open. The 'Import Options' section is highlighted with a black box and a blue border. The 'Input File' section shows the raw data content. The 'Data Frame' section shows a preview of the imported data frame.

Import Options

Name: dataSepsis

Encoding: Automatic

Heading: Yes

Row names: Automatic

Separator: Semicolon

Decimal: Period

Quote: Double quote ("")

Comment: None

na.strings: NaN

Strings as factors

Input File

```
HR;O2Sat;Temp;SBP;MAP;DBP;Resp;EtCO2;BaseExcess;HCO3;FiO2;pa
103;90;NaN;NaN;NaN;NaN;30;NaN;21;45;NaN;7.37;90;91;16;14;98
58;95;36.11;143;77;47;11;NaN;NaN;22;NaN;NaN;NaN;NaN;100
91;94;38.5;133;74;48;34;NaN;NaN;31;0.8;NaN;NaN;NaN;NaN;30;N
92;100;NaN;NaN;NaN;NaN;NaN;NaN;29;NaN;NaN;NaN;NaN;9
155.5;94.5;NaN;147.5;102;NaN;33;NaN;-12;13;1;7.22;36;NaN;45
73;99;36.06;100;67;49.5;16.5;NaN;-8;16;NaN;7.27;37;NaN;NaN;
NaN;NaN;NaN;NaN;NaN;NaN;NaN;0;25;NaN;7.35;48;NaN;NaN;NaN;
82;100;35.5;112;79.5;63;14;NaN;0;23;1;7.42;37;NaN;NaN;18;NaN;
89;100;NaN;141;85;57;17;NaN;1;25;NaN;7.43;37;NaN;NaN;9;NaN;
100;95;37.28;121;20;NaN;NaN;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;
95;100;NaN;89;62.33;NaN;18;NaN;NaN;22;NaN;NaN;NaN;NaN;8;19;
86;96;38;111;66;49;17;NaN;1;27;NaN;7.39;45;95;NaN;16;NaN;NaN;
88;100;36.3;99;66;52;16;NaN;-3;20;1;7.35;39;NaN;NaN;14;NaN;
116;97;38.28;200;108;90;24;NaN;6;NaN;0.7;7.51;39;NaN;NaN;NaN;
```

Data Frame

V1	V2	V3	V4	V5	V6	V7	V8	V9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess
103	90	NaN	NaN	NaN	NaN	30	NaN	21
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-12
73	99	36.06	100	67	49.5	16.5	NaN	-6
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-1

Raw Data File Preview

To Be Imported Data File Preview



Import Dataset Name

The screenshot shows the RStudio interface with the following components:

- Left Panel:** A terminal window titled "Imported Dataset Name" containing R version 4.1.0 startup information.
- Middle Panel:** An "Import Dataset" dialog box. The "Name" field is set to "dataSepsis". Other settings include:
 - Encoding: Automatic
 - Heading: No
 - Row names: Automatic
 - Separator: Semicolon
 - Decimal: Period
 - Quote: Double quote (")
 - Comment: None
 - na.strings: NaN
 - Strings as factors
- Right Panel:** A preview of the "Data Frame" with columns labeled v1 through v9. The first few rows of data are as follows:

v1	v2	v3	v4	v5	v6	v7	v8	v9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess
103	90	NaN	NaN	NaN	NaN	30	NaN	21
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-1

Encoding Options

The screenshot shows the RStudio interface with the 'Import Dataset' dialog box open. The 'Name' field is set to 'dataSepsis'. The 'Encoding' dropdown is set to 'Automatic', which is highlighted with a blue selection bar. The 'Input File' pane displays the first few lines of a CSV file named 'sepsis_train.csv'. The 'Data Frame' pane shows the structure of the imported dataset with columns labeled v1 through v8.

Encoding Options
(How the Data is Encoded)

R is a collaborative pr
Type 'contributors()' f
'citation()' on how to
Type 'demo()' for some
'help.start()' for an H
Type 'q()' to quit R.

> |

Import Dataset

Name: dataSepsis

Encoding: Automatic

Heading: Automatic

Row names: 437

Separator: 850

Decimal: 852

Quote: 852

Comment: 855

na.strings: 857

Strings as:

860

861

862

863

865

866

869

ANSI_X3.4-1968

ANSI_X3.4-1986

ASCII

Input File

Data Frame

v1 v2 v3 v4 v5 v6 v7 v8 v9

v1	v2	v3	v4	v5	v6	v7	v8	v9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess
103	90	NaN	NaN	NaN	30	NaN	2.1	14.98
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1.5
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-3

An Introduction to R The R Language Definition

Heading, Row Names

The screenshot shows the RStudio interface with the following components:

- Console Tab:** Displays R session output including the R version (4.1.0), system information (x86_64-w64-mingw32), and a welcome message.
- Import Dataset Dialog:** A modal window titled "Import Dataset".
 - Name:** Set to "dataSepsis".
 - Encoding:** Set to "Automatic".
 - Heading:** Set to "Yes" (radio button selected).
 - Row names:** Set to "Automatic".
 - Separator:** Set to "Automatic" (highlighted in blue).
 - Decimal:** Set to "Use first column".
 - Quote:** Set to "Use numbers".
 - Comment:** Set to "NaN".
 - na.strings:** An unchecked checkbox.
- Data Frame View:** A preview of the imported dataset "dataSepsis". The columns are labeled V1 through V9. The data includes various numerical values and some NaN entries.
- Project View:** Shows a list of files and a search bar.
- Help View:** Shows links to IDE Support, Community Forum, Cheat Sheets, Tip of the Day, Packages, and Products.

Heading
(Header Row – Yes or
No)

How to Name the
Rows

Type 'demo()' for some
'help.start()' for an H
Type 'q()' to quit R.

Separator

The screenshot shows the RStudio interface with the 'Import Dataset' dialog open. The 'Separator' dropdown is highlighted with a blue arrow pointing from a tooltip. The tooltip text is:

Separator
(What separates the fields in the data?)

The 'Separator' dropdown menu includes options: Semicolon (selected), Whitespace, Comma, Tab, and na.strings.

The 'Input File' section displays the first few lines of the dataset:

```
HR;O2Sat;Temp;SBP;MAP;DBP;Resp;EtCO2;BaseExcess;HCO3;FiO2;p
103;90;NaN;NaN;NaN;30;NaN;21;45;NaN;7.37;90;91;16;14;98
58;95;36.11;143;77;47;11;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;100
91;94;38.5;133;74;48;34;NaN;NaN;31;0.8;NaN;NaN;NaN;NaN;30;N
92;100;NaN;NaN;NaN;NaN;NaN;29;NaN;NaN;NaN;NaN;NaN;NaN;9
155.5;94.5;NaN;147.5;102;NaN;33;NaN;-12;13;1;7.22;36;NaN;45
73;99;36.06;100;67;49.5;16.5;NaN;-8;16;NaN;7.27;37;NaN;NaN;
NaN;NaN;NaN;NaN;NaN;NaN;0.25;NaN;7.35;48;NaN;NaN;NaN;NaN;
82;100;35.5;112;79.5;63;14;NaN;0;23;1;7.42;37;NaN;NaN;18;Na
89;100;NaN;141;85;57;17;NaN;1;25;NaN;7.43;37;NaN;NaN;9;NaN;
100;95;37.28;121;20;NaN;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;NaN;
95;100;NaN;89;62.33;NaN;18;NaN;NaN;22;NaN;NaN;NaN;NaN;8;19;
86;96;38;111;66;49;17;NaN;1;27;NaN;7.39;45;95;NaN;16;NaN;Na
88;100;36.3;99;66;52;16;NaN;-3;20;1;7.35;39;NaN;NaN;14;NaN;
116;97;38.28;200;108;90;24;NaN;6;NaN;0.7;7.51;39;NaN;NaN;Na
```

The 'Data Frame' section shows the first few rows of the imported dataset:

V1	V2	V3	V4	V5	V6	V7	V8	V9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess
103	90	NaN	NaN	NaN	NaN	30	NaN	21
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-1



Decimal

Decimal
(What symbol represents the decimal in numbers?)

The screenshot shows the RStudio interface with the 'Import Dataset' dialog box open. The 'Decimal' dropdown is set to 'Period' (highlighted with a blue arrow), while other options like 'Comma' and 'Space' are visible. The 'Input File' preview shows a series of numbers separated by commas. The 'Data Frame' below displays the actual data with columns labeled V1 through V9.

V1	V2	V3	V4	V5	V6	V7	V8	V9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess
103	90	NaN	NaN	NaN	NaN	30	NaN	21
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-3

Quote

Quote
(How many quotation marks enclose a string)

The screenshot shows the RStudio interface. On the left, the Console tab displays R version 4.1.0 (2021-05-18) running on an x86_64-w64-mingw32 platform. The output includes a quote about quotation marks and some code related to help.start(). On the right, the Import Dataset dialog is open, showing settings for importing 'dataSepsis'. The 'Quote' dropdown is set to 'Double quote ("') and is highlighted with a blue arrow. The 'Input File' section shows the first few lines of the CSV data. Below the dialog, a Data Frame preview shows columns V1 through Vs with various numerical values.

R version 4.1.0 (2021-05-18) Copyright (C) 2021 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32

Help start() for an introduction to R. Type 'q()' to quit R.

Import Dataset

Name: dataSepsis

Encoding: Automatic

Heading: Yes (selected)

Row names: Automatic

Separator: Semicolon

Decimal: Period

Quote: Double quote ("") (selected)

Comment: Double quote ("")

na.strings: None

Data Frame

V1	V2	V3	V4	V5	V6	V7	V8	V9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess
103	90	NaN	NaN	NaN	NaN	30	NaN	21
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-1

Comment

Comment
(What symbol denotes a comment)

The screenshot shows the RStudio interface. On the left, the R console window displays the R environment:

```
R version 4.1.0 (2021-05-18) -- "Software Suggestion"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

Type ? for help, and %? for help on a specific function.
Type demo() for some basic examples.
Type help.start() for an HTML browser help start.
Type 'q()' to quit R.
```

A blue arrow points from the text "laborative contributors()' f" in the console to the "Comment" dropdown in the "Import Dataset" dialog. The "Comment" dropdown is set to "None".

The "Import Dataset" dialog is open, showing the following settings:

- Name: dataSepsis
- Input File: A large text area containing a CSV file with many rows of data, starting with HR, O2Sat, Temp, SBP, MAP, DBP, Resp, EtCO2, BaseExcess, HCO3, FIO2, pO2, and various NaN values.
- Encoding: Automatic
- Heading: No
- Row names: Automatic
- Separator: Semicolon
- Decimal: Period
- Quote: Double quote ("")
- Comment: None
- na.strings: None
- Strings as: #

The "Data Frame" section below the dialog shows the first few rows of the imported data:

v1	v2	v3	v4	v5	v6	v7	v8	v9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess
103	90	NaN	NaN	NaN	NaN	30	NaN	21
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-1

On the right side of the RStudio interface, there is a sidebar with links to various resources:

- Studio
- IDE Support
- Community Forum
- Cheat Sheets
- Tip of the Day
- Packages
- Products

At the bottom of the RStudio window, there are navigation links: "An Introduction to R" and "The R Language Definition".



Na.strings and strings as factors

The screenshot shows the RStudio interface with the following components:

- Console Tab:** Displays R version 4.1.0 (2021-04-22) running on a Mac OS X 10.15.7 system.
- Import Dataset Dialog:** A modal window titled "Import Dataset".
 - Name:** dataSepsis
 - Encoding:** Automatic
 - Heading:** No
 - Row names:** Automatic
 - Separator:** Semicolon
 - Decimal:** Period
 - Quote:** Double quote (")
 - Comment:** None
 - na.strings:** NaN
 - Strings as factors:**
- Data Frame View:** A table titled "Data Frame" showing the first 20 rows of the dataset.

v1	v2	v3	v4	v5	v6	v7	v8	v9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtcO2	BaseExcess
103	90	NaN	NaN	NaN	NaN	30	NaN	21
58	95	36.11	143	77	47	11	NaN	NaN
91	94	38.5	133	74	48	34	NaN	NaN
92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	NaN
95	100	NaN	89	62.33	NaN	18	NaN	NaN
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-3
- Help Sidebar:** Shows links to IDE Support, Community Forum, Cheat Sheets, Tip of the Day, Packages, and Products.

na.strings
(What string
denotes a null or
n/a value)

Strings as factors
(Checking this box
will make R
recognize any
character strings as
factors or character
variables)

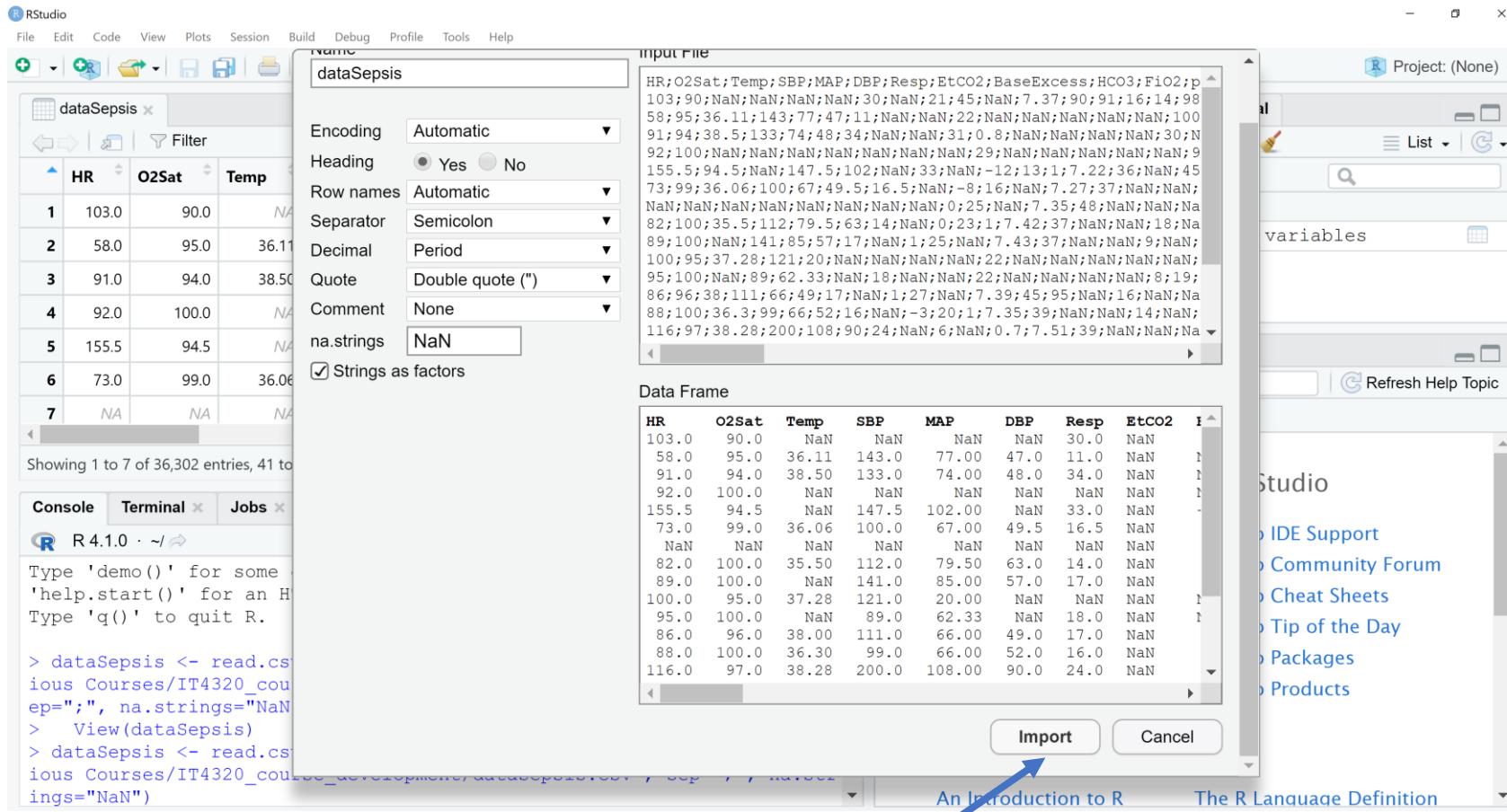
The Options Displayed are Those Required to Successfully Import the sepsis dataset.

The screenshot shows the RStudio interface with the following details:

- Console Tab:** Displays R version 4.1.0 (2021-05-18) Copyright (C) 2021 The R Foundation for Statistical Computing. It also shows the command: > dataSepsis <- read.csv("sepsis_train.csv", na.strings="NaN", header=TRUE, sep=";") > View(dataSepsis)
- Import Dataset Dialog:** Shows the configuration for importing the sepsis dataset.
 - Name:** dataSepsis
 - Encoding:** Automatic
 - Heading:** Yes
 - Row names:** Automatic
 - Separator:** Semicolon
 - Decimal:** Period
 - Quote:** Double quote ("")
 - Comment:** None
 - na.strings:** NaN
 - Strings as factors
- Data Frame Preview:** A preview of the imported data frame with columns: HR, O2Sat, Temp, SBP, MAP, DBP, Resp, EtCO2, I. The data starts with:

HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	I
103.0	90.0	NaN	NaN	NaN	NaN	30.0	NaN	1
58.0	95.0	36.11	143.0	77.00	47.0	11.0	NaN	1
91.0	94.0	38.50	133.0	74.00	48.0	34.0	NaN	1
155.5	94.5	NaN	147.5	102.00	NaN	33.0	NaN	1
73.0	99.0	36.06	100.0	67.00	49.5	16.5	NaN	1
82.0	100.0	35.50	112.0	79.50	63.0	14.0	NaN	1
89.0	100.0	NaN	141.0	85.00	57.0	17.0	NaN	1
100.0	95.0	37.28	121.0	20.00	NaN	NaN	NaN	1
95.0	100.0	NaN	89.0	62.33	NaN	18.0	NaN	1
86.0	96.0	38.00	111.0	66.00	49.0	17.0	NaN	1
88.0	100.0	36.30	99.0	66.00	52.0	16.0	NaN	1
116.0	97.0	38.28	200.0	108.00	90.0	24.0	NaN	1
- Project Explorer:** Shows the Project is (None).
- Help:** Provides links to IDE Support, Community Forum, Cheat Sheets, Tip of the Day, Packages, and Products.

Scroll Down and Click “Import” to Complete Import Process



You May Verify Successful Upload On the Following Screen

The screenshot shows the RStudio interface with three main sections highlighted by blue brackets:

- Imported Data:** A data grid titled "dataSepsis" showing 7 rows of data from a CSV file. The columns include HR, O2Sat, Temp, SBP, MAP, DBP, Resp, EtCO2, BaseExcess, and HCC.
- Dataset Summary:** The "Environment" tab of the Global Environment pane, which lists the "dataSepsis" object as containing 36,302 observations and 41 variables.
- Logfile Indicating Options in the Import Process:** The Console pane showing R code used to import the CSV file, including the command `read.csv` with options like `header=FALSE`, `sep=";", and `na.strings="NaN"`. It also shows the `View(dataSepsis)` command.

Logfile
Indicating
Options in
the Import
Process



More Options for Importing Data Into R Studio

<https://support.rstudio.com/hc/en-us/articles/218611977-Importing-Data-with-the-RStudio-IDE>

